ITC 3/55
Information Technology and Control
Vol. 55 / No. 3/ 2026
pp. 1019-1036
DOI 10.5755/j01.itc.55.3.42479

YOLO-DMSA: Improved Low-Light Object Detection via Adaptive Feature Extraction and Attention Fusion

Received 2025/08/06 | Accepted after revision 2025/12/24

HOW TO CITE: Qiang, J., Jia, Q., Zhao, L., Xu, T., Ouyang, Z. (2026) YOLO-DMSA: Improved Low-Light Object Detection via Adaptive Feature Extraction and Attention Fusion. *Information Technology and Control*, 55(3), 1019-1034. https://doi.org/10.5755/j01.itc.55.3.42479

YOLO-DMSA: Improved Low-Light Object Detection via Adaptive Feature Extraction and Attention Fusion

Jun Qiang, Qihang Jia, Li Zhao, Tianqi Xu, Zhaomeng Ouyang
School of Computer and Information, Anhui Polytechnic University, Wuhu,241000, China

Corresponding author: chiang_j@ahpu.edu.cn

In low-light environments, owing to the restrictions of illumination conditions, images are affected by problems such as contrast loss, noise interference, and color distortion, which notably increase the challenge of object detection. To address the aforementioned problems, this paper proposes an improved method called YOLO-DMSA. First, to address the issue of capturing global information in low-light images, a Dual-Mode Adaptive Feature Extraction module is designed, enabling the efficient extraction of the global features of objects. Second, aiming at the problems of low attention to key regions in low-light images and significant background noise interference, Manhattan-SLA module is constructed to focus attention on key regions and suppress noise interference. Finally, considering the characteristics of large object-scale variations in low-light environments, YOLO-DMSA upgrades the detection head to ASFFHead, which realizes the dynamic fusion of feature maps at diverse scales, thereby strengthening the detection robustness of multi-scale objects. Experimental outcomes on the DarkFace and ExDark datasets show that, compared with the baseline algorithm, YOLO-DMSA attains a 1.8% and 3.5% increase in mAP@0.5, respectively, verifying that the proposed method exhibits favorable detection performance. The code is available at https://github.com/Jiaqihang1/YOLO-DMSA.

KEYWORDS: Object detection, YOLOv11n, Deep learning, Low-light, Attention mechanism.

1. Introduction

Detecting objects in poorly illuminated conditions poses a major difficulty for computer vision systems, yet holds substantial practical importance across multiple domains including self-driving vehicles [29], surveillance systems [1], and drone-based inspections [35]. The multiple challenges in object detection under low-light environments essentially stem from image quality degradation caused by insufficient illumination, and these issues overlap with each other to further amplify the detection difficulty. In practical scenarios, insufficient illumination directly leads to the loss of image contrast: the differences in grayscale and brightness between targets and the background are significantly reduced, and originally clear contour boundaries become blurred. For instance, in nighttime surveillance, the distinguishability between pedestrians and dark walls drops sharply; in autonomous driving scenarios, the contrast between vehicles and dim road surfaces is drastically diminished. Meanwhile, interference signals such as sensor thermal noise and electronic noise are significantly intensified in low-light environments, creating noise interference that further damages image details – during drone-based inspections, for example, the texture of connectors on power transmission lines is obscured by noise under low light, resulting in the loss of critical structural information. Additionally, insufficient illumination can cause color distortion: images under streetlights often exhibit a yellowish or reddish color cast, leading to inaccuracies in target classification that relies on color features. These three types of issues interact with each other: they not only blur the boundaries between targets and the background but also cause massive loss of key image details, making it difficult for detection algorithms to extract effective features and greatly increasing the challenges in target localization and classification. However, traditional detection algorithms often suffer from severe performance degradation in low-light environments, leading to reduced equipment efficiency and potential safety risks. Therefore, efficient and robust low-light object detection methods are of practical significance for enhancing the adaptability of intelligent devices to harsh environments [7].

Current approaches to object detection under low-light conditions can be broadly divided into two categories. The first category of methods adopts a strategy that combines image enhancement algorithms and object detection. Such methods first employ image enhancement algorithms, such as SCINet [18], Retinex-Net [28], and Zero-DCE [8], to improve image quality and offset the decline in detection accuracy caused by insufficient illumination. The refined images are subsequently input into detection algorithms for object identification. For example, Xu et al. [31] developed an enhanced Faster-RCNN-based [22] framework specifically optimized for vehicle detection in nighttime scenarios. This approach integrates deformable convolutional networks (DCN) [3] into the Faster-RCNN architecture while employing Side-Aware Boundary Localization (SABL) [26] as an alternative to conventional bounding box regression, thereby simultaneously enhancing detection precision in low-light conditions and decreasing computational overhead. Wang et al. [27] introduced an innovative object recognition framework for low-light scenarios by integrating Deep Convolutional Generative Adversarial Networks (DCGAN) [19] with Faster-RCNN. Their methodology involves feeding nocturnal images into the DCGAN architecture, where adversarial learning between the generator and discriminator modules produces synthetic daytime-like representations of target scenes. The integrated system subsequently employs Faster-RCNN to boost detection accuracy through two complementary strategies: cross-layer feature integration and pyramidal ROI alignment, specifically designed for nighttime vision challenges. Ren et al. introduced YOLO-AS [21], a novel object detection framework specifically designed for complex dark environments. This architecture effectively integrates the Zero-DCE image enhancement module with YOLOv7 [25], demonstrating significant improvements in both detection accuracy and multi-scale feature representation capability in low-light conditions. Liu et al. proposed Dark-YOLO [16], which is based on YOLOv8 [23]. Through adaptive image enhancement, improved spatial feature pyramid, dynamic feature extraction, and dimensional attention mechanism, it demonstrates enhanced detection performance in low-light environments.

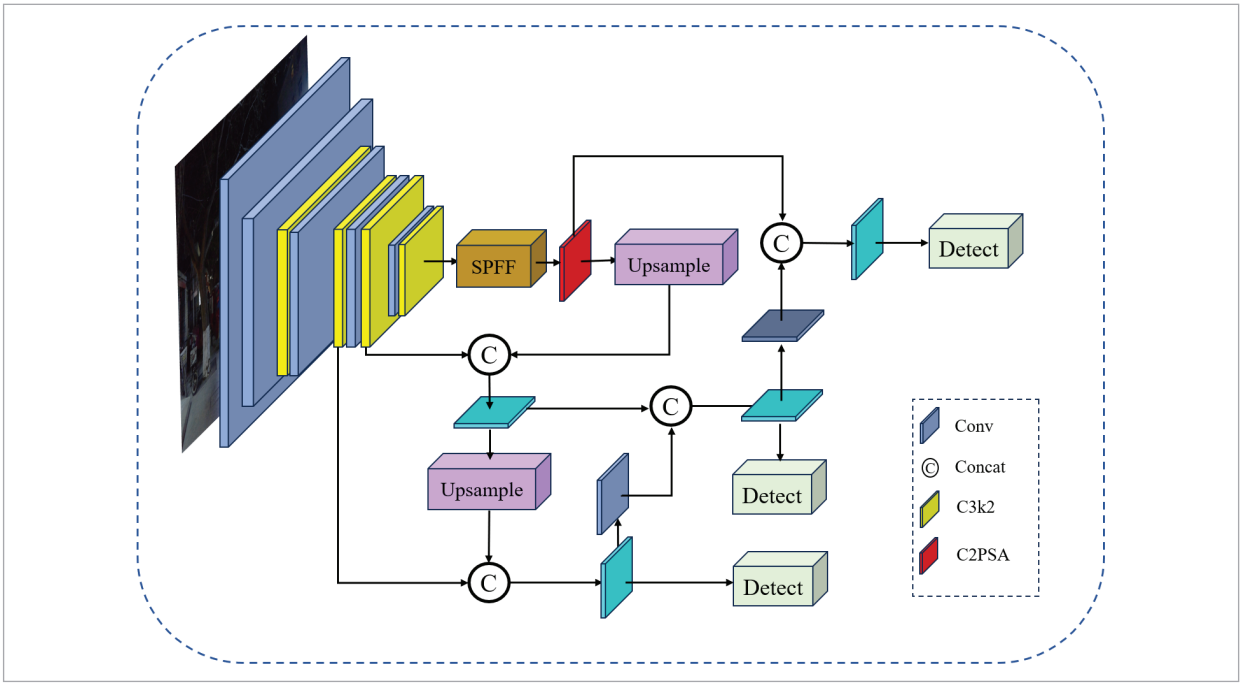
However, the existing paradigm of sequentially executing image enhancement and object detection processes exhibits notable limitations. At the computational level, an architecture in which the two modules are executed independently fails to share and reuse computing resources, resulting in severe computational redundancy. At the performance collaboration level, owing to the lack of a joint optimization mechanism, the image enhancement process does not fully consider the feature requirements of the detection network. Consequently, although the enhanced images exhibit better visual quality for human observers, they often fail to meet the optimal input specifications required by detection networks. This incompatibility constrains the performance improvement of low-light object detection systems.

The second category of methods focuses on improving the object detection algorithm. These methods directly optimize the detection model to maintain high accuracy under low-light conditions without relying on additional image enhancement modules. YOLO [20] represents a widely adopted approach within contemporary object detection frameworks. This architecture demonstrates superior computational efficiency compared to alternative detection methodologies and has been widely used for low-light object recognition. For example, Yin et al. developed PE-YOLO [33], an enhanced framework integrating Pyramid Enhancement Network (PENet) architecture with YOLOv3 [6]. This approach employs Laplacian pyramid decomposition to progressively refine image details and low-frequency components across multiple scales, thereby boosting detection accuracy in low-light environments. Liu et al. developed IA-YOLO [15], incorporating a novel Differentiable Image Processing (DIP) component into the detection framework. Its parameters are dynamically predicted by a small CNN, enabling the joint optimization of image enhancement and object detection, thus achieving favorable detection performance under both foggy and low-light conditions. Han et al. proposed 3L-YOLO[10] based on YOLOv8n, it reduces computational costs without requiring an image enhancement module. Wu et al. proposed FuDA-YOLO [30], an advanced framework that enhances detection accuracy in low-light environments through domain adaptation with multi-scale feature integration.

The second category of methods, by virtue of the characteristic of directly optimizing the detection model, exhibits strong adaptability and high efficiency in low-light scenarios. Furthermore, Chen et al. proposed the Collaborative Compensative Transformer Network (CCTNet) [2], which takes image patches and superpixels as visual tokens to establish long-range dependencies, and leverages collaborative and compensative fusion transformers to associate global context, localize object boundaries, and thus achieve effective feature enhancement for ambiguous boundaries and indistinct features in low-light scenes. Kim et al. presented the Inverse Saliency Pyramid Reconstruction Network (InSPyReNet) [13] for high-resolution salient object detection, which adopts a strict saliency map image pyramid integrated with fusion technology to resolve effective receptive field (ERF) discrepancies, preserves object details without high-resolution training data, and mitigates detail loss and noise interference in low-light high-resolution image detection. However, current approaches have the potential for optimization in terms of both computational efficiency and performance synergy. Breaking through the bottlenecks of computational redundancy and enhancement adaptability while maintaining efficient detection has become an urgent problem to be solved.

Compared with other methods, YOLOv11n [11] has architectural advantages in feature learning and multi-scale detection. Its backbone network adopts the C3k2 module instead of the traditional C2f module, enhancing feature expression capability through a dual-branch design. The original PAN neck structure was replaced with the C3k2 module to enhance multi-scale feature fusion capability. Additionally, the decoupled head architecture was integrated with a novel P2 detection layer, significantly improving the small-object detection performance. However, images captured in low-light environments generally suffer from issues such as amplified noise, reduced contrast, and blurred edges. The backbone network and neck design of YOLOv11 primarily focus on optimizing computational efficiency and multi-scale feature fusion, lacking robust designs tailored to address low-light degradation. The C3k2 module enhances feature learning capability through a dual-branch structure, but it does not incorporate a noise suppression mechanism. This makes feature maps highly susceptible to

Figure 1
Structure Diagram of YOLOv11 Network.



noise interference under low-light conditions, which impairs subsequent detection accuracy. Low-light environments further exacerbate the challenges of small object detection – problems such as small target pixel proportion and loss of texture information often arise. Nevertheless, the decoupled head design of YOLOv11 exhibits insufficient sensitivity to tiny features in low-light scenarios. The structure of YOLOv11 is shown in Figure 1.

In conclusion, this paper adopted YOLOv11n as the baseline architecture and conducted systematic structural enhancements to develop YOLO-DM-SA, a novel detection framework specifically optimized for low-light conditions. The proposed method maintains the lightweight characteristics of the model while significantly boosting feature extraction, attention focusing, and multi-scale fusion capabilities. Major technical innovations concentrate on three core aspects:

1 To effectively overcome the limitation of the local receptive field of ordinary convolutions, thereby enhancing the ability to perceive blurred targets and weak boundary features in low-light images, improving the global modeling effect of overall feature expression, and significantly boosting feature fusion capability while maintaining model inference efficiency, this paper designs a Dual-Mode Adaptive Feature Extraction (DMAFE) module and embeds it into the backbone network. This module replaces the original convolution module with a CMUNeXtBlock [24] that includes depthwise separable convolution and a large kernel structure, and introduces RepConv through RepBottleneck to support consistent reparameterization in both training and inference phases[4].

2 To suppress redundant interference caused by complex backgrounds in low-light images and make the model more sensitive to salient target features under insufficient illumination, thereby improving the robustness and accuracy of detection, the improved method introduces an adaptive attention fusion module, namely the Manhattan-SLA (MSLA) module, in the neck part. This module combines Manhattan Self-Attention (MaSA)[5] and Simplified Linear Attention (SLA)[9], employing trainable parameters to adaptively balance their weight outputs through dual channel-spatial enhancement of critical target responses.

3 To address the challenges of feature redundancy and interference while enhancing the detection precision for small and edge targets, the Adaptive Spatial Feature Fusion (ASFF) module [14] was integrated into the detection head, which was upgraded to ASFFHead. The ASFFHead enables spatially-adaptive weighted fusion of multi-scale feature maps, thereby improving the multi-scale object detection capability of the network.

2. Methods

2.1. DMAFE

Traditional convolution has limited flexibility in extracting multi-scale features because the fixed size of its kernels. To improve multi-scale feature representation and computational efficiency, this paper designed the DMAFE module, the architectural details of which are presented in Figure 2. When the parameter c3k=True, the DMAFE module employs C3k_Rep, which is suitable for tasks that require a larger receptive field. When c3k=False, the DMAFE module switches to Bottleneck_CMU and focuses on lightweight global feature extraction.

The DMAFE module first improves the C3k part of the C3k2 module in YOLOv11n into C3k_Rep, which introduces RepBottleneck and replaces the first-layer convolution with RepConv. Relying on its multi-branch structure design, it captures multi-scale features during the training phase and is reparameterized into a single branch during the inference phase to ensure real-time performance. The structure of C3k_Rep is shown in Figure 3.

To facilitate concurrent extraction of global and local features without compromising computational efficiency, the DMAFE module further optimizes the C3k2 module's bottleneck structure in YOLOv11n, introducing an enhanced Bottleneck_CMU. Bottleneck_CMU replaces the convolution in the standard bottleneck layer of YOLOv11n with a CMUNeXtBlock. The structure of CMUNeXtBlock is shown in Figure 4.

CMUNeXtBlock adopts large convolution kernels and depth-wise convolutions, exhibiting excellent

Figure 2
Structure Diagram of the DMAFE Module.

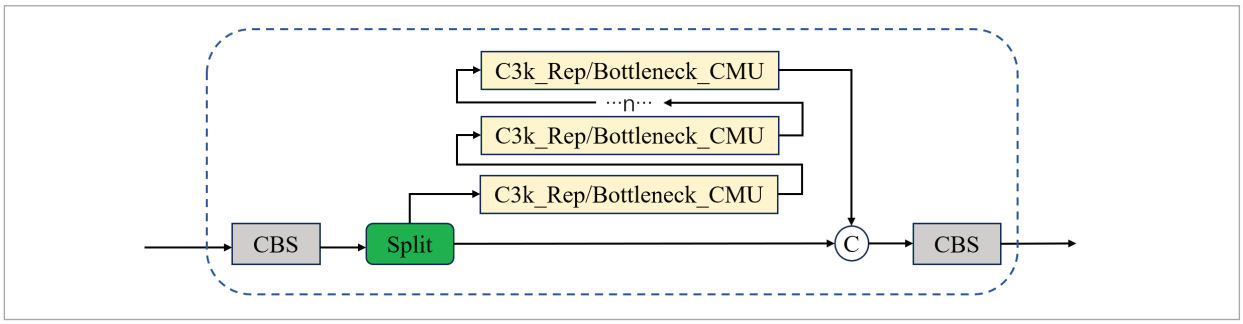


Figure 3
Structure Diagram of the C3k_Rep Module.

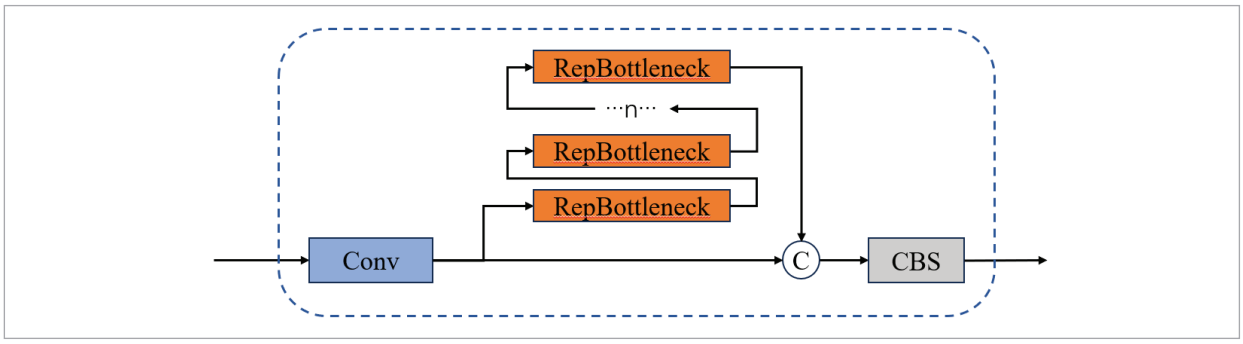
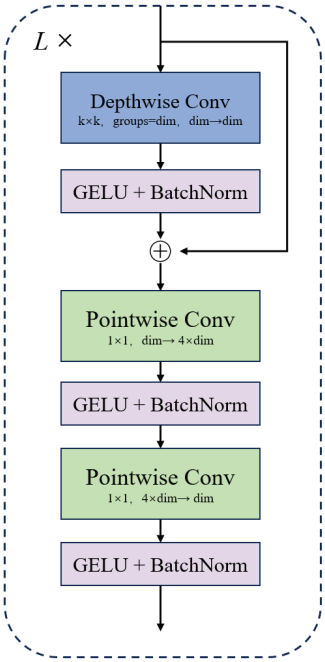


Figure 4
Structure diagram of the CMUNeXtBlock module.



performance in expanding the receptive field and capturing long-distance features. Second, it introduces depth-wise separable convolution, which decomposes standard convolution into two steps: depth-wise convolution and point-wise convolution, significantly reducing the number of parameters and computational costs, thereby lowering resource consumption. Furthermore, it expands the intermediate feature dimension through an inverted bottleneck structure, achieving deep fusion of spatial and channel information. Finally, this module applies residual connections and batch normalization techniques to ensure the stability of model training, accelerate the convergence process, and improve inference efficiency.

In the task of object detection in low-light environments, the DMAFE module can flexibly capture multi-scale features and expand the receptive field to address the problem of varying target scales. Meanwhile, it efficiently extracts both global and local features to adapt to uneven feature distributions under low-light conditions. Additionally, its lightweight design improves inference efficiency, enabling the model to complete object detection more accurately and rapidly in complex low-light scenarios.

2.2. MSLA

In object detection for low-light images, the blurring of global features and degradation of local spatial structures pose significant challenges. Under insufficient lighting conditions, visual data frequently suffer from weakened global contrast and unclear semantic details, limiting the effectiveness of standard convolutional operations in establishing long-range feature relationships. Additionally, the presence of amplified noise artifacts and degraded edge sharpness adversely affects the precise characterization of local spatial features, consequently impairing target localization and recognition accuracy. To address these issues, this paper constructs an MSLA module.

The MSLA module is an adaptive attention fusion module that combines two attention mechanisms: the MaSA and SLA modules. The MaSA module optimizes spatial information modeling for visual tasks by introducing a spatial decay matrix based on the Manhattan distance, which injects explicit spatial priors to compensate for the insufficient spatial perception of traditional self-attention. In contrast, the SLA module implements an optimized linear attention framework that employs ReLU-based kernel computation integrated with depthwise convolutional operations to strengthen local feature representation capabilities. It can achieve an approximate global attention effect with linear complexity, thereby effectively capturing long-range dependencies.

The MSLA module constructed in this paper combines these two attention mechanisms with different advantages, leveraging their strengths to improve the model's performance in object detection tasks under low-light environments, thereby forming a more powerful and robust feature extraction scheme. This module adopts learnable weight parameters to adaptively adjust the contribution ratio of the two attention mechanisms, enabling the model to automatically select the most appropriate feature extraction strategy according to the different input features. Specifically, the output of the MSLA module is obtained by the weighted summation of the output feature map O_{MasA} from MaSA and the output feature map O_{SLA} from SLA, as shown in Formula (1):

$$O_{MSLA} = \alpha \cdot O_{MasA} + (1-\alpha) \cdot O_{SLA}, \quad (1)$$

wherein, the weighting coefficient α is derived by mapping the learnable parameter α_{logit} through the sigmoid function, as shown in Formula (2):

$$\alpha=\text{sigmoid}(\alpha_{\text{logit}}) \tag{2}$$

and α_{logit} is initialized as the log-odds form corresponding to the initial proportion of the target, as shown in Formula (3):

$$\alpha_{\text{logit}}=\log(\alpha_0/(1-\alpha_0)). \tag{3}$$

After being constrained by the sigmoid function, the value range of α was limited to the interval [0, 1]. This design enables the model to adaptively optimize weight allocation during training based on features such as the blurriness of targets and noise distribution in low-light images: when the input image has severe boundary blurring, the model can increase α to enhance the spatial perception ability of MaSA; when global context information is more critical for detection, it can decrease α to highlight the global modeling advantage of SLA. Ultimately, this achieves adaptive feature extraction in complex low-light scenarios.

2.3. ASFFHead

To mitigate multi-scale feature redundancy and conflict in low-light environments while improving the detection accuracy for small and edge-degraded targets, the ASFF module was incorporated into an enhanced detection head architecture. This modification optimizes the semantic consistency between the hierarchical feature representations and target localization. The working principle of the ASFF module is illustrated in Figure 5.

During the initialization phase of the detection head, ASFFHead constructs independent ASFF sub-modules for the three feature maps of different levels output by the network. Each ASFF sub-component implements adaptive multi-scale feature fusion via a learnable weighting mechanism, allowing the network to dynamically focus on more important spatial positions and scale information. Taking ASFF-3 as an example, the formula for calculating its output feature map is shown in Formula (4):

$$Y=X^{1\to3}\otimes\alpha^3+X^{2\to3}\otimes\beta^3+X^{3\to3}\otimes\gamma^3. \tag{4}$$

In the Formula (4), Y indicates the output feature map of the ASFF-3 module. $X^{1\to3}$, $X^{2\to3}$, and $X^{3\to3}$ respectively denote the feature maps adjusted from each scale to the 3rd scale. α^3, β^3, and γ^3 are the weighting coefficients corresponding to the 1st, 2nd, and 3rd scale feature maps, respectively. These weighting coefficients are normalized to satisfy $\alpha^3+\beta^3+\gamma^3=1$. $\otimes$ denotes the element-wise multiplication operation, which involves multiplying the feature maps $X^{1\to3}$, $X^{2\to3}$, and $X^{3\to3}$ by their corresponding weighting coefficients α^3, β^3, and γ^3 element-wise. Finally, the fused feature map Y is obtained by accumulation.

Conflicts may arise between features of different scales during the feature fusion process. Small targets are prone to interference from targets of normal size, leading to a loss of information. Furthermore, excessive irrelevant information from adjacent layers may be inadvertently incorporated during fusion,

Figure 5

Working principle diagram of the ASFF module.

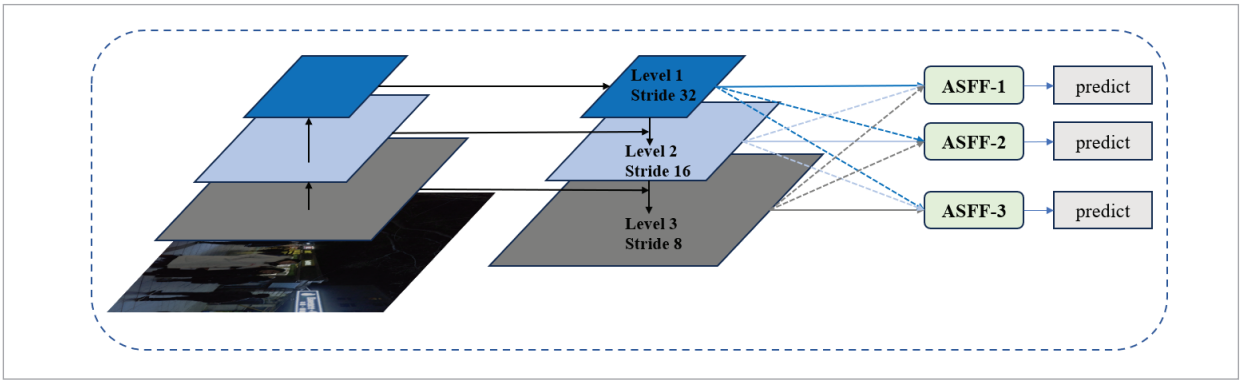
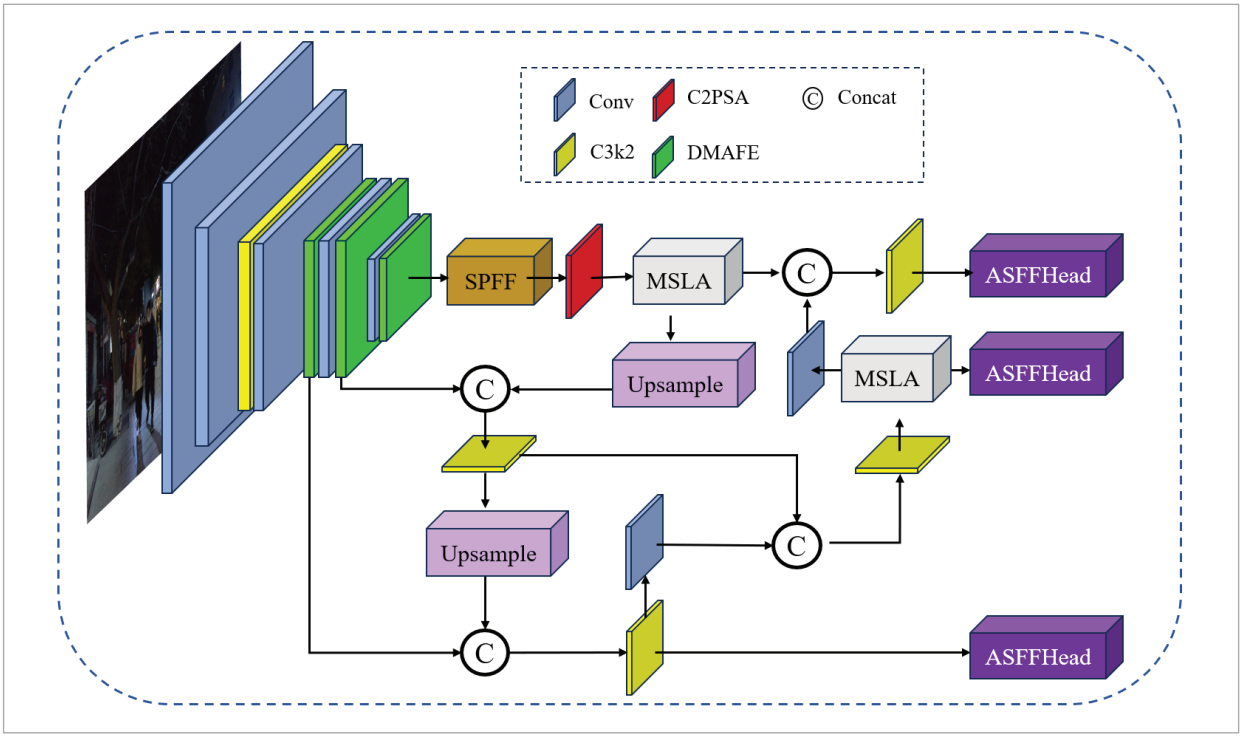


Figure 6
Structure Diagram of the YOLO-DMSA Network.



potentially leading to detection failure and compromised recognition performance. The adaptive spatial feature fusion structure can effectively classify the feature maps obtained after feature extraction by the backbone network, perform adaptive learning, filter features at different levels, and retain and combine the useful information. It not only preserves the semantic information of high-level features but also pays attention to the detailed information of low-level features, effectively improving the situation of missing target feature information and making the detection of small and multiple targets more accurate.

2.4. YOLO-DMSA

The overall framework of YOLO-DMSA inherits the three-stage architecture of the YOLO series, consisting of three parts: backbone, neck, and head. The proposed approach builds upon the YOLOv11n architecture to achieve enhanced object detection performance under low-light conditions via specialized structural modifications. Its core innovation lies in the integration of the DMAFE module, MSLA module, and ASFFHead.

In the backbone section, the improved DMAFE module replaces the original C3k2 module of YOLOv11n. While retaining the dual-branch design, it introduces lightweight global feature extraction capabilities. When c3k=True, the C3k_Rep module with a large receptive field is activated, replacing the first-layer convolution with RepConv. During the training phase, multi-scale features are captured through a multi-branch structure, and during inference, it is reparameterized into a single-branch structure to balance performance and real-time capabilities. When c3k=False, it switches to Bottleneck_CMU, replacing the convolution operations in the standard bottleneck layer. Through the collaborative design of large convolution kernels and depth-wise separable convolutions, it significantly suppresses noise interference in low-light conditions and improves the contour recognition of nighttime targets.

In the neck stage, the MSLA module is embedded, which introduces a spatial decay matrix based on Manhattan distance to inject explicit spatial priors into the attention mechanism, enabling accurate localization of blurred target boundaries under low-

light conditions. The MSLA module dynamically weights and fuses the outputs of MaSA and SLA through learnable parameters, allowing the network to automatically adjust the feature extraction strategy under different lighting conditions. While maintaining linear computational complexity, it effectively addresses the issues of global semantic ambiguity and local structural degradation in low-light images.

Finally, YOLO-DMSA improves the original detection head into ASFFHead, which constructs independent fusion paths for different feature maps respectively. Through learnable parameters, it performs adaptive weighted fusion on features of different levels, suppresses the interference of cross-layer redundant information, and strengthens the feature expression of small targets and edge-blurred targets.

YOLO-DMSA enhances its capabilities in noise suppression, multi-scale perception, and boundary localization significantly while maintaining real-time performance, achieved through three key improvements: lightweight global modeling in the backbone network (via DMAFE), robust attention fusion in the neck network (via MSLA), and adaptive feature optimization in the detection head (via ASFFHead). Figure 6 illustrates the detailed architecture of the enhanced network structure.

3. Experimental Results

3.1. Experimental Environment and Evaluation Indicators

In this study, Python 3.9 was used as the programming language, paired with the PyTorch 1.9.0 deep learning framework, and the experiments were conducted on the Windows 11 operating system. The hardware configuration included an NVIDIA GeForce RTX 3080 graphics card, an Intel (R) Core (TM) i9-10940X processor, and the CUDA 11.0 computing platform.

The default hyperparameter settings are as follows: the number of epochs is 100, the batch size is 16, the optimizer used is SGD (Stochastic Gradient Descent), the initial learning rate is 0.001, the momentum factor is set to 0.937, and the learning rate decay coefficient is 0.3.

Additionally, to ensure fair comparisons in this study's experiments, no pre-trained weights were adopted for any model—all models were trained from scratch. For boundary box regression, the loss function employs CIoU (Complete Intersection over Union), the most widely used loss function in the YOLO series.

The evaluation framework incorporates five key metrics: precision (P), recall (R), mAP@0.5, mAP@0.5:0.95, the number of parameters, floating-point operations (FLOPs) and Frames Per Second (FPS). Precision refers to the model's ability to identify relevant targets, and the percentage of accurately detected results among all detections is called precision. Recall is an important indicator to measure the model's ability to identify positive examples. Mean average precision (mAP) is the average of the detection accuracies of all categories; mAP@0.5 refers to the Mean AP (average precision) of all categories when the Intersection over Union (IoU) is set to 0.5; mAP@0.5:0.95 is the average mAP at different IoU thresholds (from 0.5 to 0.95 with a step of 0.05); FLOPs represent the computational complexity of the model. FPS measures the model's inference speed, defined as the number of images processed by the model per second, and serves as a core metric for evaluating real-time detection performance. The corresponding mathematical formulations are presented below:

$$\text{Preccision}=\frac{\text{TP}}{\text{TP}+FP} \tag{5}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+FN} \tag{6}$$

$$AP_i=\int_0^1 P_i(R_i)\text{dR}_i \tag{7}$$

$$mAP@0.5=\frac{1}{N}\sum_{i=1}^{N} AP_i \tag{8}$$

$$mAP@0.5:0.95=\frac{1}{T}\sum_{t=0.5}^{0.95} AP_t \tag{9}$$

$$IoU=\frac{intersection}{union}. \tag{10}$$

Figure 7
Label distribution information of the ExDark dataset.

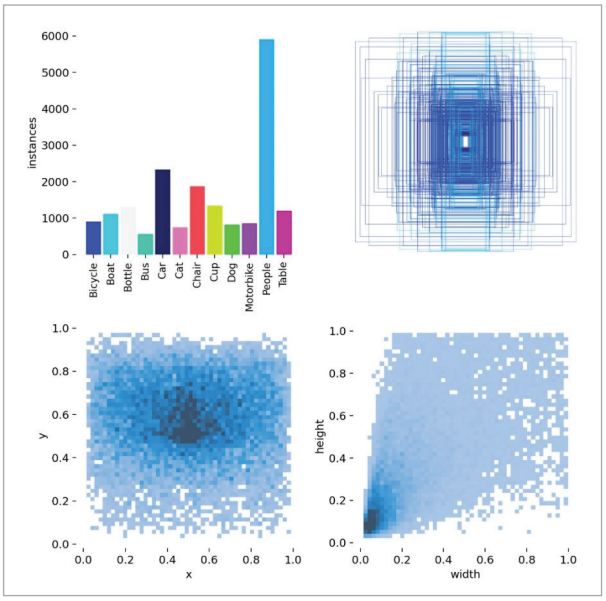
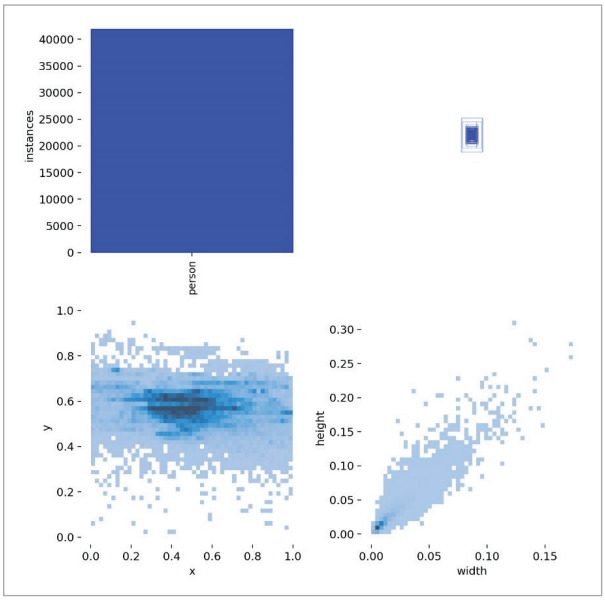


Figure 8
Label distribution information of the DarkFace dataset.



In Formulas (5, 6), TP (True Positive) refers to the number of cases where the prediction is positive and the true value is also positive; FP (False Positive) is the number of cases where the prediction is positive but the actual value is negative; FN (False Negative) represents the number of cases incorrectly detected as negative samples. In Formulas (7, 8), N is the number of categories, and AP_i is the average precision of the i-th category. In Formula (9), T is the number of IoU thresholds, and AP_t is the average precision at the IoU threshold t. In Formula (10), IoU denotes the overlap between the coordinates of the predicted bounding box and the ground-truth box, where Intersection is the area of the overlapping part of the two regions, and Union is the total area of the two regions.

3.2. Experimental Dataset

The experimental evaluation employs two specialized datasets: ExDark [17] for general low-light object detection tasks and DarkFace [32] for low-light face detection scenarios.

ExDark comprises 7,363 images captured under 10 illumination conditions (from extremely to moderately low light), annotated with 12 object categories (bicycle, boat, bottle, bus, car, cat, chair, cup, dog, motorcycle, person, table). It was split into training (5,884 images), validation (741 images), and test (738 images) sets at an 8:1:1 ratio. The dataset features diverse natural light sources to ensure generalization, with label statistics (Figure 7) showing: training set sample counts per category (top-left); target bounding box size/quantity distribution (top-right, revealing concentration in specific aspect ratio ranges); bounding box center positional distribution (bottom-left, reflecting target dominance in image centers, consistent with common shooting habits); and target aspect ratio distribution (bottom-right, indicating wide but medium-size-focused target scales).

DarkFace contains 6,000 images with a single "Person" category, exhibiting more extreme low-light conditions than ExDark for increased detection challenge. It was split into training (5,000 images), validation (500 images), and test (500 images) sets. Figure 8 show most face targets are small (width: 0–0.15, height: 0–0.3), leading to higher detection difficulty than ExDark.

3.3. Experimental Results of Random Seeds

This section investigates the influence of random seeds on model performance. In deep learning, random seeds regulate stochastic operations, which may introduce minor performance variability. To validate the training robustness of the proposed model, we perform an ablation study on random

Table 1
Experimental Results of Random Seeds on the DarkFace dataset.

Seed	P	R	mAP@0.5	mAP@0.5:0.95
0	0.689	**0.415**	**0.463**	0.198
42	0.677	0.413	0.458	0.197
123	0.688	0.412	0.457	0.198
456	**0.694**	0.409	0.460	**0.199**
789	0.685	0.414	0.456	0.196
1000	0.682	0.411	0.458	0.198

seeds: six distinct seeds (0, 42, 123, 456, 789, 1000) are tested, with the model trained independently under identical experimental configurations.

Experimental findings demonstrate minimal performance variability across seeds: P spans 0.677 to 0.694, R falls within 0.409–0.415, mAP@0.5 varies from 0.456 to 0.463, and mAP@0.5:0.95 remains within 0.196–0.199. This consistency suggests the model exhibits low sensitivity to stochastic variations, thereby validating its strong training stability.

For subsequent experiments, the default random seed 0 is adopted based on two considerations: (1) Among evaluated seeds, seed 0 yields a relatively superior mAP@0.5 score (0.463); (2) Employing the default seed 0 conforms to conventional experimental practices, enhancing result reproducibility and facilitating subsequent comparative analyses.

3.4. Results of Comparative Experiments

To evaluates the object detection performance of the proposed YOLO-DMSA method in low-light environments, this experiment compares the capability of different methods on the ExDark low-light dataset, including YOLOv3-tiny, YOLOv5n, YOLOv7-tiny, YOLOv10n, YOLOv11n, YOLOv12n[12], RT-DETR-ResNet50 (Real-Time Detection Transformer)[34] and the proposed YOLO-DMSA method. The comparative experimental results are shown in Table 2. Meanwhile, partial experimental results relevant to YOLO-based models are visualized as line charts, thereby facilitating a more intuitive comparison of how design optimizations impact key metrics across iterations. The visualization results are illustrated in Figure 9.

Figure 9
Visualization results of comparative experiments on the ExDark dataset.

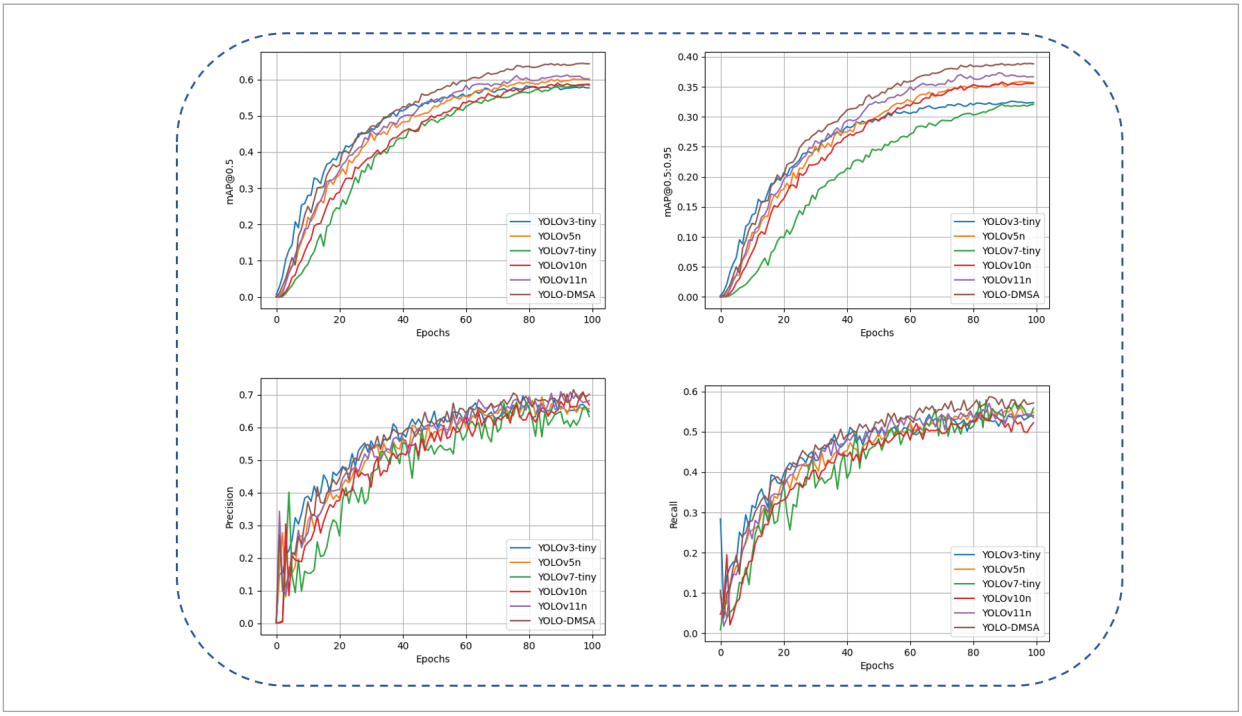


Table 2
Results of comparative experiments on the ExDark dataset.

Method	P	R	mAP@0.5	mAP@0.5:0.95	Params(M)	GFLOPs	FPS
YOLOv3-tiny	0.664	0.526	0.576	0.326	9.53	14.3	**163**
YOLOv5n	0.662	0.541	0.600	0.358	**2.19**	**5.8**	93
YOLOv7-tiny	0.634	0.559	0.587	0.321	6.01	13.2	101
YOLOv10n	0.655	0.526	0.590	0.358	2.71	8.3	85
YOLOv11n	0.685	0.541	0.609	0.368	2.59	6.3	92
YOLOv12n	0.660	0.539	0.594	0.364	2.55	6.3	80
RT-DETR-ResNet50	0.701	0.567	0.627	0.388	42.76	126.7	34
YOLO-DMSA	**0.702**	**0.586**	**0.644**	**0.389**	7.10	9.6	71

As can be seen from Table 2 and Figure 9, YOLO-DMSA achieves a balanced improvement in all metrics compared to other methods. Compared with the baseline network YOLOv11n, precision of YOLO-DMSA increases from 68.5% to 70.2%, with an improvement of 1.7%; recall rate rises from 54.1% to 58.6%, an increase of 4.5%; mAP@0.5 goes up from 60.9% to 64.4%, a boost of 3.5%; mAP@0.5:0.95 increases from 36.8% to 38.9%, an improvement of 2.1%. Although the computational complexity increases from 6.3 GFLOPs to 9.6 GFLOPs and the number of parameters also rises slightly, the model still meets the requirements for real-time detection, which indicates that the improvement strat-

Figure 10
Object detection results on ExDark dataset.



egy is effective. Compared with methods such as YOLOv3-tiny, YOLOv5n, YOLOv7-tiny, YOLOv10n, YOLOv12n and RT-DETR-ResNet50, the mAP@0.5 of YOLO-DMSA increases by 6.8%, 4.4%, 5.7%, 5.4%, 5.0% and 1.7%respectively, and the mAP@0.5:0.95 increases by 6.3%, 3.1%, 6.8%, 3.1%, 2.5% and 0.1% respectively. Experimental results demonstrate that YOLO-DMSA achieves robust object detection performance in low-light environments while satisfying real-time processing requirements, establishing it as a viable solution for detection tasks in low-light environments.

Figure 10 presents the prediction results of YOLO-DMSA in comparison with methods that perform well on the ExDark dataset, such as YOLOv5n, YOLOv10n, and YOLOv11n. It can be seen from the figure that, compared with other methods, YOLO-DMSA effectively mitigates issues like missed detections and false detections in object detection under low-light environments.

To better evaluate the generalization and robustness of the YOLO-DMSA method for object detection in low-light environments, this paper also conducts comparative experiments on the more challenging DarkFace low-light face detection dataset. The experimental results are shown in Table 3. Compared with the ExDark dataset, the mAP indicators of all methods generally decrease on the DarkFace dataset, indicating that the task difficulty of the DarkFace dataset is significantly higher, especially in the strict indicator mAP@0.5:0.95. Partial experimental results are also visualized as line charts, and the visualization results are shown in Figure 11.

The DarkFace evaluation demonstrates YOLO-DMSA's superior performance with a 48.1% mAP@0.5, exceeding the YOLOv11n baseline by 1.8%. Precision and recall metrics show respective improvements of 1.9% (68.9%→70.8%) and 1.4% (41.5%→42.9%). At the same time, it has improvements in various indicators compared with other methods. Notably, RT-DETR-ResNet50 fails to deliver satisfactory results in this more challenging DarkFace environment: its bulky 42.76M parameter structure is prone to introducing noise interference in low-light, low-contrast scenes, and the excessive 126.7 GFLOPs not only slows its inference speed to a mere

Figure 11

Visualization results of the comparative experiments on the DarkFace dataset.

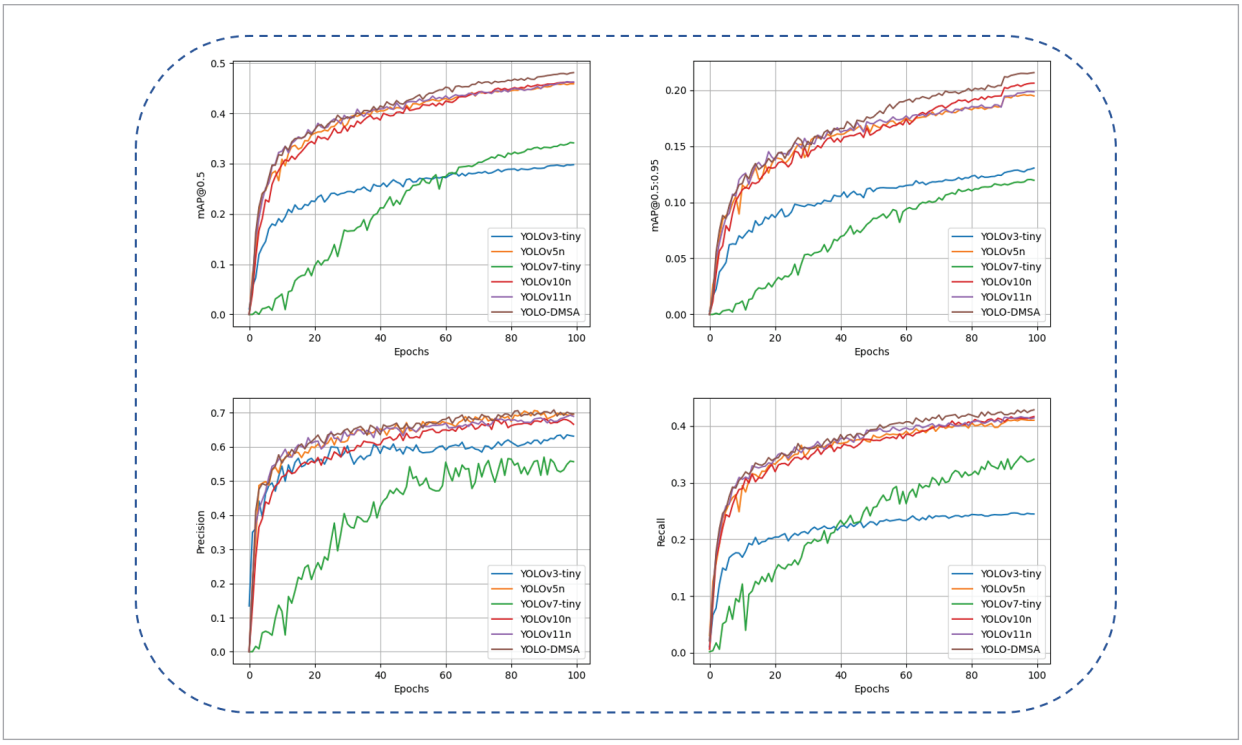


Figure 12
Object detection results on DarkFace dataset.



Table 3
Results of comparative experiments on the DarkFace dataset.

Method	P	R	mAP@0.5	mAP@0.5:0.95	Params(M)	GFLOPs	FPS
YOLOv3-tiny	0.632	0.245	0.298	0.130	9.53	14.3	**190**
YOLOv5n	0.697	0.411	0.459	0.195	**2.19**	**5.8**	98
YOLOv7-tiny	0.557	0.342	0.341	0.120	6.01	13.2	144
YOLOv10n	0.667	0.417	0.462	0.206	2.71	8.3	83
YOLOv11n	0.689	0.415	0.463	0.198	2.59	6.3	98
YOLOv12n	0.682	0.405	0.446	0.190	2.55	6.3	74
RT-DETR-ResNet50	0.567	0.389	0.408	0.160	42.76	126.7	33
YOLO-DMSA	**0.708**	**0.429**	**0.481**	**0.216**	7.10	9.6	67

33 FPS but also restricts its capture of subtle weak-light object features, leading to less competitive detection metrics here. Figure 12 also shows the prediction results of YOLO-DMSA in comparison with methods that perform well on the DarkFace dataset, such as YOLOv5n, YOLOv10n, and YOLOv11n. The findings of the experiment demonstrate that the proposed YOLO-DMSA approach shows remarkable generalization ability and robustness in object detection tasks within low-light conditions.

3.5. Results of Ablation Experiments

For verifying the enhancement effect of each module on low-light object detection, ablation experiments are performed on the ExDark dataset. The specific experimental results are shown in Table 4.

According to the experimental results, when the MSLA module is added alone, compared with the original YOLOv11n method, the mAP@0.5 increases from 60.9% to 62.4%, an improvement of 1.5%, and the computational complexity rises from 6.3 GFLOPs to 7.2 GFLOPs. When the DMAFE module is added alone, the mAP@0.5 reaches 62.0%, which is 1.1% higher than that of the original method, and the computational complexity slightly increases to 6.6 GFLOPs. This module can improve the object detection accuracy in low-light environments to a certain extent by optimizing feature fusion. When ASFFHead is introduced alone, the mAP@0.5 reaches 61.4%, indicating that its adaptive feature fusion mechanism can also bring gains to detection accuracy in low-light scenarios.

Furthermore, the effects of combining multiple modules were tested. The mAP@0.5 of the YOLOv11n + MSLA + DMAFE method reached 62.9% with a computational complexity of 7.4 GFLOPs, indicating that the synergy of multiple modules strengthened feature processing to a certain extent and improved accuracy. The mAP@0.5 of the YOLOv11n + MSLA + ASFFHead method remained at 62.4%, suggesting that this combination did not achieve efficient gains. The mAP@0.5 of the YOLOv11n + DMAFE + ASFFHead method reached 62.3%, which was an improvement compared to introducing the two modules alone.

In contrast, the YOLO-DMSA method, which integrates all effective modules, attained a mAP@0.5 of

Table 4
Results of ablation experiments on the ExDark dataset.

Method	P	R	mAP@0.5	mAP@0.5:0.95	GFLOPs
YOLOv11n	0.685	0.541	0.609	0.368	**6.3**
YOLOv11n+MSLA	0.684	0.555	0.624	0.383	7.2
YOLOv11n+DMAFE	0.694	0.538	0.620	0.376	6.6
YOLOv11n+ASFFHead	0.687	0.546	0.614	0.366	8.5
YOLOv11n+MSLA+DMAFE	0.629	0.576	0.629	0.380	7.4
YOLOv11n+MSLA+ASFFHead	**0.718**	0.543	0.624	0.381	8.9
YOLOv11n+DMAFE+ASFFHead	0.708	0.538	0.623	0.377	8.8
YOLO-DMSA	0.702	**0.586**	**0.644**	**0.389**	9.6

Table 5
Results of ablation experiments on the DarkFace dataset.

Method	P	R	mAP@0.5	mAP@0.5:0.95	GFLOPs
YOLOv11n	0.689	0.415	0.463	0.198	**6.3**
YOLOv11n+MSLA	0.670	0.424	0.469	0.208	7.2
YOLOv11n+DMAFE	0.671	0.418	0.458	0.195	6.6
YOLOv11n+ASFFHead	0.695	0.428	0.474	0.211	8.5
YOLOv11n+MSLA+DMAFE	0.698	0.427	0.473	0.211	7.4
YOLOv11n+MSLA+ASFFHead	0.697	0.429	0.477	0.215	8.9
YOLOv11n+DMAFE+ASFFHead	0.684	0.425	0.476	0.214	8.8
YOLO-DMSA	**0.708**	**0.429**	**0.481**	**0.216**	9.6

64.4% and a mAP@0.5:0.95 of 38.9%, with precision and recall reaching 70.2% and 58.6% respectively. Although the computational complexity slightly increased to 9.6 GFLOPs, the accuracy improved significantly. Experimental confirms the efficacy of module synergy in enhancing target recognition and localization capabilities under low-light conditions, demonstrating significant performance gains in challenging visual environments.

Similarly, to verify the performance of each module in more challenging low-light environments, ablation experiments were conducted on the DarkFace low-light face detection dataset, and the experimental results are shown in Table 5.

According to the experimental results, compared with the original YOLOv11n method, after adding different modules, the precision, recall, mAP@0.5, and mAP@0.5:0.95 have generally improved to varying degrees. Although GFLOPs have increased, they still meet the requirements of real-time object detection. This indicates that these modules also have strong generalization and robustness in object detection tasks under low-light environments.

In summary, each module plays a unique role in object detection for low-light images: The DMAFE module replaces the original convolution. By leveraging depth-wise separable convolutions, large convolution kernel structures, and reparameterization, it breaks through the local limitations of ordinary convolutions. It enhances the perception of blurred targets and weak boundary features in low-light conditions, thereby improving global modeling and feature fusion capabilities. After the MSLA module is integrated into the neck part, it fuses two attention mechanisms and dynamically adjusts weights. This suppresses redundant information in complex backgrounds, strengthens the feature response of salient targets under low light, and improves detection robustness and accuracy. Integrating the ASFF module into the detection head enables spatial weighted fusion of multi-scale features. It alleviates feature redundancy and conflicts, and optimizes the detection of small targets and edge targets. When each module collaboratively forms YOLO-DMSA, it outperforms the base method, single-module methods, and partial module combination methods in key indicators such as mAP@0.5 and mAP@0.5:0.95. This proves that the improvement idea of significantly enhancing feature extraction, attention focusing, and multi-scale fusion capabilities while maintaining lightweight through structural optimization is effective. The modules complement each other's advantages, making the model more adaptable to the needs of object detection in low-light scenarios. It provides a better solution for the accurate recognition and localization of targets in low-light images, verifying the effectiveness and rationality of the YOLO-DMSA method.

4. Conclusion

To tackle issues such as insufficient contrast, loss of details, and noise interference in object detection for low-light images, this paper proposes an improved method based on YOLOv11n, namely YOLO-DMSA. The network is optimized from three aspects: receptive field expansion, feature focusing, and multi-scale fusion, so as to enhance the model's recognition capability and robustness in low-light scenarios. Firstly, the designed DMAFE module combines depth-wise separable convolution with large convolution kernels, which expands the receptive field and strengthens the ability to model global information. Secondly, the constructed MSLA module can dynamically adjust the information weight between global context and local details, focusing on key regions while suppressing noise, and maintains linear computational complexity. Finally, the ASFFHead module incorporates trainable spatial weighting coefficients to adaptively integrate cross-scale features, significantly enhancing multi-scale object recognition performance. Ablation experiments and comparative experiments conducted on the DarkFace and ExDark datasets verify the effectiveness and robustness of the proposed method in various low-light scenarios. Without significantly increasing the computational burden, this method improves the accuracy and stability of low-light object detection, showing promising application prospects. Despite the aforementioned achievements, this study still has certain limitations: for instance, its performance in extremely low-light environments and complex backgrounds can be further improved, the real-time efficiency of the model on resource-constrained devices needs additional optimization,

and the generalization ability of the model across different types of low-light scenarios remains to be enhanced. Future research directions will therefore focus on several key aspects: developing more effective end-to-end joint optimization methods for low-light image enhancement and object detection, exploring dynamic network structures based on attention mechanisms, conducting experimental validation in more real-world scenarios.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62401011, in part by the Middle-Aged and Young Teachers Cultivate Action of Anhui Province under Grant YQZD2023041, and in part by the Anhui Polytechnic University industry-university-research cooperation project under Grant HX-2023-12-124.

References

1. Ai, S., Kwon, J. Extreme Low-Light Image Enhancement for Surveillance Cameras Using Attention U-Net. Sensors, 2020, 20(2), 495. https://doi.org/10.3390/s20020495
2. Chen, J., Zhang, H., Gong, M., Gao, Z. Collaborative Compensative Transformer Network for Salient Object Detection. Pattern Recognition, 2024, 154, 110600. https://doi.org/10.1016/j.patcog.2024.110600
3. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y. Deformable Convolutional Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, 764-773. https://doi.org/10.1109/ICCV.2017.89
4. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J. RepVGG: Making VGG-Style ConvNets Great Again. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 13728-13737. https://doi.org/10.1109/CVPR46437.2021.01352
5. Fan, Q., Huang, H., Chen, M., Liu, H., He, R. RMT: Retentive Networks Meet Vision Transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, 5641-5651. https://doi.org/10.1109/CVPR52733.2024.00539
6. Farhadi, A., Redmon, J. YOLOv3: An Incremental Improvement. Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2018, 1-6.
7. Ghari, B., Tourani, A., Shahbahrami, A., Gaydadjiev, G. Pedestrian Detection in Low-Light Conditions: A Comprehensive Survey. Image and Vision Computing, 2024, 105106. https://doi.org/10.1016/j.imavis.2024.105106
8. Guo, C., Li, C., Guo, J., Loy, C. C., Hou, J., Kwong, S., Cong, R. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 1777-1786. https://doi.org/10.1109/CVPR42600.2020.00185
9. Guo, J., Chen, X., Tang, Y., Li, K., Wang, Y. SLAB: Efficient Transformers with Simplified Linear Attention and Progressive Re-Parameterized Batch Normalization. arXiv Preprint arXiv:2405.11582, 2024.
10. Han, Z., Yue, Z., Liu, L. 3L-YOLO: A Lightweight Low-Light Object Detection Algorithm. Applied Sciences, 2025, 15(1), 90. https://doi.org/10.3390/app15010090
11. Khanam, R., Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. arXiv Preprint arXiv:2410.17725, 2024.
12. Khanam, R., Hussain, M. A Review of YOLOv12: Attention-Based Enhancements vs. Previous Versions. arXiv Preprint arXiv:2504.11995, 2025.
13. Kim, T., Kim, K., Lee, J., Kim, J. Revisiting Image Pyramid Structure for High Resolution Salient Object Detection. Proceedings of the Asian Conference on Computer Vision (ACCV), 2022, 108-124. https://doi.org/10.1007/978-3-031-26293-7_16
14. Liu, S., Huang, D., Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. arXiv Preprint arXiv:1911.09516, 2019.
15. Liu, W., Ren, G., Yu, R., Guo, S., Zhu, J., Zhang, L. Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2), 1792-1800. https://doi.org/10.1609/aaai.v36i2.20072
16. Liu, Y., Li, S., Zhou, L., Liu, H., Li, Z. Dark-YOLO: A Low-Light Object Detection Algorithm Integrating Multiple Attention Mechanisms. Applied Sciences, 2025, 15(9), 5170. https://doi.org/10.3390/app15095170
17. Loh, Y. P., Chan, C. S. Getting to Know Low-Light Images with the Exclusively Dark Dataset. Computer Vision and Image Understanding, 2019, 178, 30-42. https://doi.org/10.1016/j.cviu.2018.10.010

18. Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z. Toward Fast, Flexible, and Robust Low-Light Image Enhancement. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 5637-5646. https://doi.org/10.1109/CVPR52688.2022.00555

19. Radford, A., Metz, L., Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. Computer Science, 2015.

20. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

21. Ren, B., Xu, Z., Zhao, J., Hao, R., Zhang, J. Complex Dark Environment-Oriented Object Detection Method Based on YOLO-AS. Scientific Reports, 2025, 15(1), 21873. https://doi.org/10.1038/s41598-025-07348-0

22. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

23. Sohan, M., Sai Ram, T., Rami Reddy, C. V. A Review on YOLOv8 and Its Advancements. Proceedings of the International Conference on Data Intelligence and Cognitive Informatics (DICI), 2024, 529-545. https://doi.org/10.1007/978-981-99-7962-2_39

24. Tang, F., Ding, J., Quan, Q., Wang, L., Ning, C., Zhou, S. K. CMUNext: An Efficient Medical Image Segmentation Network Based on Large Kernel and Skip Fusion. Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI), 2024, 1-5. https://doi.org/10.1109/ISBI56570.2024.10635609

25. Wang, C. Y., Bochkovskiy, A., Liao, H. Y. M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, 7464-7475. https://doi.org/10.1109/CVPR52729.2023.00721

26. Wang, J., Zhang, W., Cao, Y., Chen, K., Pang, J., Gong, T., Shi, J., Loy, C. C., Lin, D. Side-Aware Boundary Localization for More Precise Object Detection. Proceedings of the European Conference on Computer Vision (ECCV), 2020, 403-419. https://doi.org/10.1007/978-3-030-58548-8_24

27. Wang, K., Liu, M. Z. Object Recognition at Night Scene Based on DCGAN and Faster R-CNN. IEEE Access, 2020, 8, 193168-193182. https://doi.org/10.1109/ACCESS.2020.3032981

28. Wei, C., Wang, W., Yang, W., Liu, J. Deep Retinex Decomposition for Low-Light Enhancement. Proceedings of the British Machine Vision Conference (BMVC), 2018, 3-6.

29. Wei, J., As'arry, A., Rezali, K. A. M., Yusoff, M. Z., Ma, H., Zhang, K. A Review of YOLO Algorithm and Its Applications in Autonomous Driving Object Detection. IEEE Access, 2025, 13, 93688-93711. https://doi.org/10.1109/ACCESS.2025.3573376

30. Wu, J., Liu, H., Wang, C. FuDA-YOLO: Low-Light Object Detection Using Multi-Scale Fusion Domain Adaptation. Proceedings of the 4th International Symposium on Computer Applications and Information Technology (ISCAIT), 2025, 30-35. https://doi.org/10.1109/ISCAIT64916.2025.11010400

31. Xu, Y., Chu, K., Zhang, J. Nighttime Vehicle Detection Algorithm Based on Improved Faster-RCNN. IEEE Access, 2023, 12, 19299-19306. https://doi.org/10.1109/ACCESS.2023.3347791

32. Yang, W., Yuan, Y., Ren, W., Liu, J., Scheirer, W. J., Wang, Z. Advancing Image Understanding in Poor Visibility Environments: A Collective Benchmark Study. IEEE Transactions on Image Processing, 2020, 29, 5737-5752. https://doi.org/10.1109/TIP.2020.2981922

33. Yin, X., Yu, Z., Fei, Z., Lv, W., Gao, X. PE-YOLO: Pyramid Enhancement Network for Dark Object Detection. Proceedings of the International Conference on Artificial Neural Networks (ICANN), 2023, 163-174. https://doi.org/10.1007/978-3-031-44195-0_14

34. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, 16965-16974. https://doi.org/10.1109/CVPR52733.2024.01605

35. Zhong, H., Zhang, Y., Shi, Z., Zhang, Y., Zhao, L. PS-YOLO: A Lighter and Faster Network for UAV Object Detection. Remote Sensing, 2025, 17(9), 1641. https://doi.org/10.3390/rs17091641

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).